

Statistique et aide à la décision

Session 5

David Causeur
Agrocampus Ouest
IRMAR CNRS UMR 6625

Plan

- 1 Prédiction
- 2 Performance de prédiction
- 3 Validation croisée

Classification



The image shows a screenshot of a Futura Santé website article. The page features a navigation bar with the Futura Santé logo and menu items: Explorer, Vidéos, Photos, Experts, Forum, Codes Promo, and a search icon. The main content area has a large, stylized background image of a coronavirus particle with red spikes. The article title is "Coronavirus : les tests sérologiques sont-ils fiables ?". Below the title, there is a category tag "ACTUALITÉ" and a sub-category "Classé sous : CORONAVIRUS, EPIDEMIE, TEST SÉROLOGIQUE". At the bottom, there is a white box containing a small portrait of Julie Kern, her name "Julie Kern", her title "Rédactrice scientifique", and the publication date "Publié le 02/05/2020".

≡ **FUTURA SANTÉ**

Explorer Vidéos Photos Experts Forum Codes Promo 🔍

SANTÉ

Coronavirus : les tests sérologiques sont-ils fiables ?

ACTUALITÉ ↕ Classé sous : CORONAVIRUS, EPIDEMIE, TEST SÉROLOGIQUE

 **Julie Kern**
Rédactrice scientifique

Publié le 02/05/2020

Classification

On cherche à prédire $Y_0 \in \{y_1, \dots, y_K\}$ pour un individu pour lequel les variables explicatives $X = (X_1, \dots, X_p)'$ prennent les valeurs $x_0 = (x_{01}, \dots, x_{0p})'$

Lorsque la variable réponse est catégorielle, prédire peut signifier :

- Estimer la probabilité *a posteriori* $\mathbb{P}(Y_0 = y_k \mid X_0 = x_0)$,
- Donner une valeur $\hat{Y}_0 = y_k$ connaissant x_0

▶ `predict(mod, type)`

▶  R Studio

Classification

La règle de décision conduisant à prédire la valeur de Y_0 pour un individu pour lequel $X_0 = x_0$ est appelée règle de **classification**.

La **règle de classification de Bayes** consiste à donner à Y_0 la valeur y_k de plus grande probabilité $\mathbb{P}(Y_0 = y_k \mid X_0 = x_0)$.



Plan

- 1 Prédiction
- 2 Performance de prédiction
- 3 Validation croisée

Erreurs de prédiction

Suite de l'article de Julie Kern, futura-sciences.com, 2 mai 2020

Une sensibilité encore trop faible

Une [étude scientifique prépubliée](#) a passé au crible neuf tests immunologiques développés au Danemark. Trois sont des tests immuno-enzymatiques Elisa (*Enzyme-Linked Immuno Assay*) réalisés par un [technicien de laboratoire](#). Les six autres sont des tests appelés « *point of care* » (POC), qui donnent des résultats rapides et sont réalisables hors des laboratoires, à l'image du kit NG Biotech.

Les trois tests Elisa ont une sensibilité située entre 67 et 93 % et une spécificité entre 93 à 100 %. De leur côté, les POC démontrent une sensibilité située entre 80 et 93 % et une spécificité entre 80 et 100 %. Mais ils n'ont été testés que sur une trentaine de personnes, ce qui est insuffisant pour les homologuer. Il faudrait les éprouver sur des milliers de personnes atteintes du [Covid-19](#) et sur autant de personnes saines.

Erreurs de prédiction

Lorsque $Y = \pm 1$ est une réponse catégorielle à deux groupes, les individus pour lesquels $\hat{Y} = +1$ sont dits **positifs** et ceux pour lesquels $\hat{Y} = -1$ sont dits **négatifs**.

- Taux de vrais positifs (sensibilité) :

$$\text{TPR} = \frac{\#\{i = 1, \dots, n, \hat{Y}_i = +1, Y_i = +1\}}{\#\{i = 1, \dots, n, Y_i = +1\}}.$$

- Taux de vrais négatifs (spécificité) :

$$\text{TNR} = \frac{\#\{i = 1, \dots, n, \hat{Y}_i = -1, Y_i = -1\}}{\#\{i = 1, \dots, n, Y_i = -1\}}.$$

Erreurs de prédiction

Supposons que le test d'une maladie d'incidence $p = \mathbb{P}(Y = +1) = 0.0001$ garantisse $TPR = TNR = 0.9$

Probabilité qu'un individu positif soit malade :

$$\begin{aligned}\mathbb{P}(Y = +1 \mid \hat{Y} = +1) &= \mathbb{P}(\hat{Y} = +1 \mid Y = +1) \frac{\mathbb{P}(Y = +1)}{\mathbb{P}(\hat{Y} = +1)}, \\ &= TPR \times \frac{p}{\mathbb{P}(\hat{Y} = +1)},\end{aligned}$$

où

$$\begin{aligned}\mathbb{P}(\hat{Y} = +1) &= \mathbb{P}(\hat{Y} = +1 \mid Y = +1)p + \mathbb{P}(\hat{Y} = +1 \mid Y = -1)(1 - p), \\ &= TPR \times p + (1 - TNR) \times (1 - p), \\ &= 0.10008.\end{aligned}$$

Erreurs de prédiction

Supposons que le test d'une maladie d'incidence $p = \mathbb{P}(Y = +1) = 0.0001$ garantisse $\text{TPR} = \text{TNR} = 0.9$

Probabilité qu'un individu positif soit malade :

$$\begin{aligned}\mathbb{P}(Y = +1 \mid \hat{Y} = +1) &= \mathbb{P}(\hat{Y} = +1 \mid Y = +1) \frac{\mathbb{P}(Y = +1)}{\mathbb{P}(\hat{Y} = +1)}, \\ &= \text{TPR} \times \frac{p}{\mathbb{P}(\hat{Y} = +1)}, \\ &= 0.0009\end{aligned}$$

Conclusion : une bonne sensibilité et une bonne spécificité **ne garantissent pas** une bonne performance de prédiction !

Erreurs de prédiction

- Précision (Positive Predictive Value) :

$$\text{PPV} = \frac{\#\{i = 1, \dots, n, \hat{Y}_i = +1, Y_i = +1\}}{\#\{i = 1, \dots, n, \hat{Y}_i = +1\}}.$$

- Negative Predictive Value :

$$\text{NPV} = \frac{\#\{i = 1, \dots, n, \hat{Y}_i = -1, Y_i = -1\}}{\#\{i = 1, \dots, n, \hat{Y}_i = -1\}}.$$



Erreurs de prédiction

Compromis TPR-TNR par le choix d'un seuil de décision

$$\hat{Y}_0 = +1 \text{ si } \hat{P}(Y = +1 \mid X_0 = x_0) \geq t$$

La stratégie du choix de t dépend du compromis recherché

▶ Package ROCR

▶  Studio®

Plan

- 1 Prédiction
- 2 Performance de prédiction
- 3 Validation croisée**

Evaluation d'une performance de classification

Objectif : s'assurer que $\hat{Y}_0 \approx Y_0$?

Attention : l'évaluation à partir de $(\hat{Y}_i, Y_i)_{i=1, \dots, n}$ est optimiste.

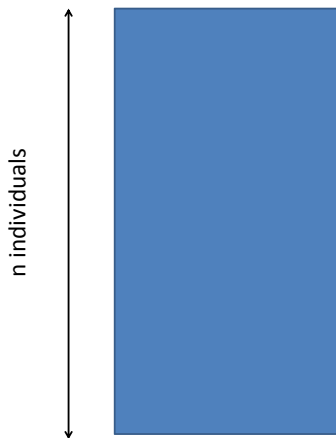
Une procédure d'évaluation à partir de (\hat{Y}_{-i}, Y_i) où la règle de prédiction ayant conduit à \hat{Y}_{-i} n'implique pas Y_i , est dite de **validation croisée**.

Recommandation : la règle de prédiction ajustée sur un **échantillon d'apprentissage** doit être évaluée par application sur un **échantillon test**, séparé de l'échantillon d'apprentissage.



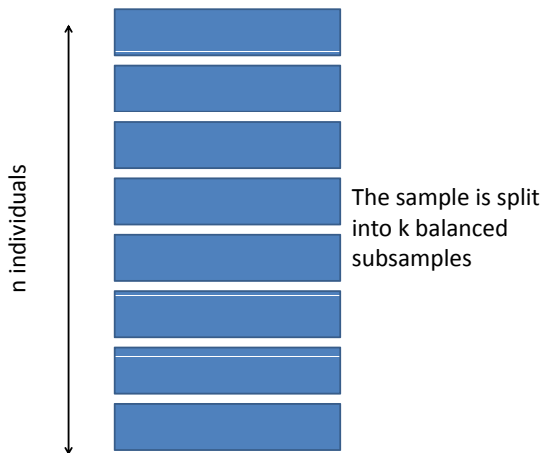
Evaluation d'une performance de classification

Si n est petit, procédure de validation croisée à K segments



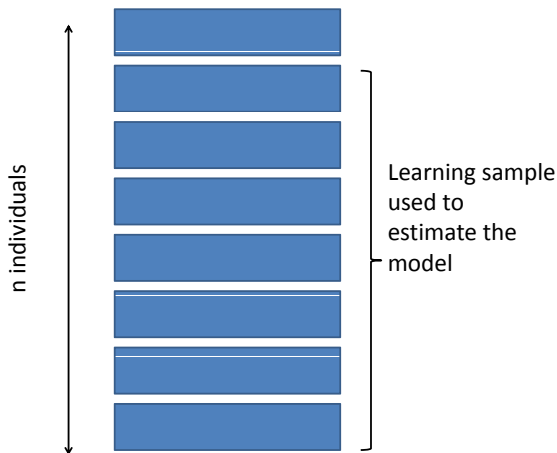
Evaluation d'une performance de classification

Si n est petit, procédure de validation croisée à K segments



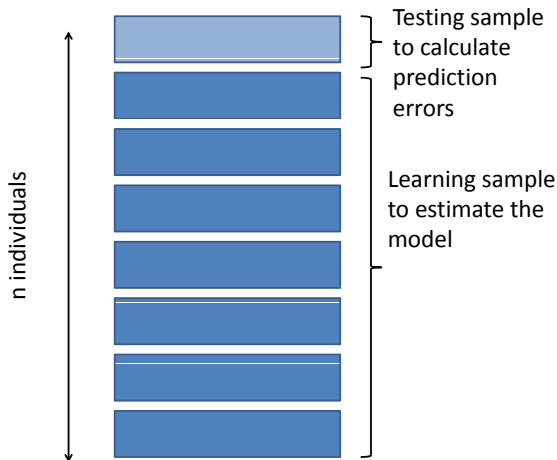
Evaluation d'une performance de classification

Si n est petit, procédure de validation croisée à K segments



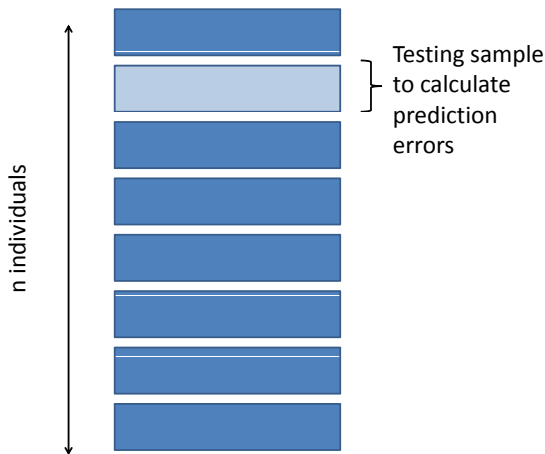
Evaluation d'une performance de classification

Si n est petit, procédure de validation croisée à K segments



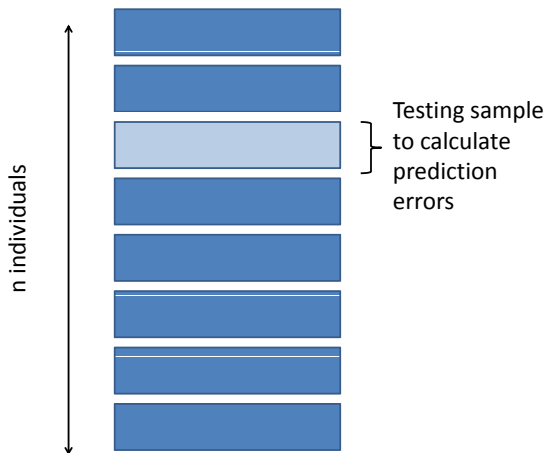
Evaluation d'une performance de classification

Si n est petit, procédure de validation croisée à K segments



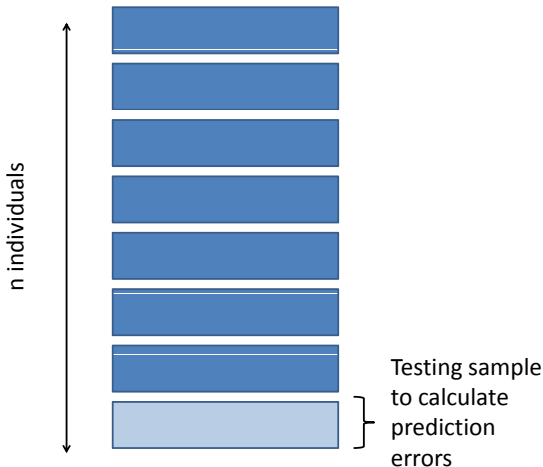
Evaluation d'une performance de classification

Si n est petit, procédure de validation croisée à K segments



Evaluation d'une performance de classification

Si n est petit, procédure de validation croisée à K segments



Evaluation d'une performance de classification

Choix de K :

- Selon le temps de calcul pour l'ajustement de la règle de classification, $K = 3$ et $K = 10$ sont souvent choisis.
- Si n est petit, $K = n$ peut-être recommandé : **leave-one-out cross-validation**.

